Lecture 2

10-701 Spring 2019

These materials are a more or less a complete summary: more details will be discussed in class and in the assigned readings.

1 Last Lecture

Last lecture:

- We estimated the probability θ that a thumbtack/coin falls on heads using data (a set of IID flips).
- This problem can be considered an example of parametric density estimation:
  - we assume the data follows a specific known distribution with unknown parameters. (here, a Bernoulli distribution with parameter θ).
  - we fit the parameters to the data. (here, find θ)

  One can also perform *non*-parametric density estimation, for example by computing a histogram or using kernel density estimation.
- To estimate the parameter θ we used the MLE estimator:

  $$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

  You can read more about this in **[KM] Chapter 6** and **[HTF] 8.2**.
- We also performed bayesian estimation and obtained the MAP estimate:

  $$\theta_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(D|\theta)P(\theta)$$

  You can read more about this in **[KM] Chapter 5** and **[HTF] 8.3**.
- We used a Bernoulli likelihood and a Beta prior, and this resulted in a Beta posterior.

  When the prior and the posterior are from the same distribution family (e.g. Beta), we call the prior and the posterior *conjugate distributions*. We also say that the prior distribution family is a *conjugate prior* for the likelihood function.

  We saw that the Beta distribution is a conjugate prior for a Bernoulli likelihood function. A family of likelihood functions (e.g. Normal Distribution) can have multiple conjugate priors (e.g. Normal, Gaussian...).

2 Probability Review

Here we quickly review some of the basic probability concepts that are required for the course. For more details refer to **[KM] Chapter 2**, **[CB] Appendix B** and the online lectures referred to in Lecture 1's slides.

Let the PDF of x be $f(x)$.

CDF

$$P(x \leq a) = \int_{-\infty}^{a} f(x)dx$$

I.e. the PDF:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

.

- $$\lim_{x \to -\infty} F(x) = 0$$
- $$\lim_{x \to +\infty} F(x) = 1$$
- $$F(b) \geq F(a) \quad \forall b \geq a$$

Expectations

$$E[x] = \int x f(x)dx$$

$E[x]$ is the mean of x.

$$E[g(x)] = \int g(x)f(x)dx$$

Variance

$$var(x) = E[(x - E[x])^2]$$
$$var(x) = E[x^2] - E[x^2]$$

$$\begin{aligned} var(x) &= E[(x - E[x])^2] \\ &= E[x^2 - 2xE[x] - E[x]^2] \\ &= E[x^2] - 2E[xE[x]] - E[E[x]^2] \\ &= E[x^2] - 2E[x]E[x] - E[x]^2 \\ &= E[x^2] - E[x]^2 \end{aligned}$$

The mean and variance of x can be obtained from the moment generating function:

$$M_x(t) = E[e^{tx}]$$

- $E[x^n] = M^{(n)}(0) = \frac{\partial^n M_x}{\partial t^n}|_{t=0}$
- $E[x] = M'(0)$
- $E[x^2] = M''(0)$
- $var(x) = E[x^2] - E[x] = M''(0) - M'(0)^2$

Joint distribution

$$P((x, y) \in A) = \int\int_A f(x, y)dxdy$$

Marginal distribution

$$f(x) = \int f(x, y)dy$$

Conditionals

$$f(x|y) = \frac{f(x, y)}{f(y)}$$

Chain Rule

$$f(x, y) = f(x|y)f(y)$$

Bayes Rule

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

In some cases, we do not have to evaluate $f(y)$ and we can use $f(x|y) \propto f(y|x)f(x)$. In other cases we have to compute $f(y)$ explicitly:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

Some distributions

Binomial

$$x \sim \text{Binomial}(p, n)$$

$$\begin{aligned} P(x = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ \binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ E[x] &= np \\ var(x) &= np(1-p) \end{aligned}$$

Uniform

$$x \sim \text{U}(p, n)$$

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \\ E[x] &= \frac{a+b}{2} \\ var(x) &= \frac{1}{12}(b-a)^2 \end{aligned}$$

Normal (Gaussian) Distribution

$$x \sim (\mu, \sigma^2)$$

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ E[x] &= \mu \\ var(x) &= \sigma^2 \end{aligned}$$

Central Limit Theorem: Suppose $x_1, x_2, ..x_n$ are IID samples with mean μ and variance $\sigma^2 < \infty$, and let $S_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$ be the sample mean. The variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal with mean 0 and variance σ^2 as n tends to infinity:

$$\sqrt{n}(S_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\xrightarrow{d}$ denotes convergence in distribution, which is one of the types of convergence of random variables we will discuss this semester.

Multivariate Normal (Gaussian)

$$\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$$

$$\begin{aligned}
\mathbf{x} &= (x_1, x_2, x_3 ..., x_M)^\top \quad \text{is a vector} \\
f(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \\
E[\mathbf{x}] &= \boldsymbol{\mu} = (\mu_1, \mu_1, \mu_1 ... \mu_M)^\top \\
&= (E[x_1], E[x_2], E[x_3] ..., E[x_M])^\top
\end{aligned}$$

Σ is the covariance matrix

$$\Sigma = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & . & . & . & cov(x_1, x_M) \\ cov(x_2, x_1) & var(x_2) & . & . & . & . \\ . & . & . & . & . & . \\ cov(x_M, x_1) & . & . & . & . & var(x_M) \end{bmatrix}$$

$$cov(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$$